Analysis of the stabilized supralinear network

Daniel B. Rubin[1,2], Yashar Ahmadian[1,4] & Kenneth D. Miller[1−4,*]

March 1, 2012

[1]*Center for Theoretical Neuroscience,* [1]*Dept. of Neuroscience,* [2]*Doctoral Program in Neurobiology and Behavior,* [3]*Swartz Program in Theoretical Neuroscience, and* [4]*Kavli Institute for Brain Science, College of Physicians and Surgeons, Columbia University, NY, NY 10032.*
[*] *To whom correspondence should be addressed: ken@neurotheory.columbia.edu.*

Abstract

We study a rate-model neural network composed of excitatory and inhibitory neurons in which neuronal input-output functions are power laws with a power greater than 1, as observed in primary visual cortex. This supralinear input-output function leads to supralinear summation of network responses to multiple inputs for weak inputs. We show that for stronger inputs, which would drive the excitatory subnetwork to instability, the network will dynamically stabilize provided feedback inhibition is sufficiently strong. This dynamic stabilization yields a transition from supralinear to sublinear summation of network responses to multiple inputs. We compare this to the dynamic stabilization in the "balanced network", which yields only linear behavior. We more exhaustively analyze the 2-dimensional case of 1 excitatory and 1 inhibitory population. We show that in this case dynamic stabilization will occur whenever the determinant of the weight matrix is positive and the inhibitory time constant is sufficiently small, and analyze the conditions for "supersaturation", or decrease of firing rates with increasing stimulus contrast (which represents increasing input firing rates). In work to be presented elsewhere, we show that this transition from supralinear to sublinear summation can explain a wide variety of nonlinearities in cerebral cortical processing.

Acknowledgements: D.B.R. is supported by NIH training grant T32-GM007367 to the M.D./Ph.D. training program at Columbia University. Y.A. is supported by a postdoctoral fellowship from the Kavli Institute for Brain Science at Columbia University. K.D.M. is supported by R01 EY11001 from the NEI of the NIH and by the Gatsby Charitable Foundation through the Gatsby Initiative in Brain Circuitry at Columbia University.

Contents

1 Introduction

We have recently found, in work that is yet unpublished except as abstracts (Miller and Rubin 2010, 2011, Rubin and Miller 2010, 2011) that a large set of response properties of cells in primary visual cortex (V1) and other sensory cortical areas can be understood from a very simple circuit motif. The response properties have in common a change in integration with increasing input strength, so that responses to weak inputs sum supralinearly while those to stronger inputs sum sublinearly.

One set of properties involve contextual modulation or "surround suppression". A visual sensory neuron has a classical receptive field (CRF), corresponding to the region in which appropriate visual stimuli will drive the neuron's responses. The size of the CRF does not change with input strength (Song and Li 2008). Stimuli outside the CRF can modulate responses to CRF stimuli, although they cannot drive responses, and typically are suppressive. However, the nature of the surround influence can vary with input strength (Polat et al. 1998, Sengpiel et al. 1997). A size tuning curve is obtained by centering an effective stimulus on the CRF center and studying response vs. stimulus radius. The summation field size is the stimulus size evoking peak response. This summation field size shrinks with input strength, as represented by stimulus contrast (Anderson et al. 2001, Cavanaugh et al. 2002, Sceniak et al. 1999, Shushruth et al. 2009, Tsui and Pack 2011). This means that regions of the surround are changing from facilitating to suppressing with increasing input strength.

Another set of properties involve sublinear summation of the responses to multiple stimuli: the response to two simultaneously presented stimuli can be closer to the average than the sum of the responses to the stimuli presented individually. We refer to this property as "normalization", because it is the most prominent of a set of nonlinear response properties that have been given that name (reviewed in Carandini and Heeger 2011). In at least some cases, this summation becomes supralinear when inputs are weak (Heuer and Britten 2002, Ohshiro et al. 2011). If one thinks of surround suppression as representing the response to simultaneous presentation of a center stimulus that normally by itself evokes a certain response and a surround stimulus that normally by itself evokes zero response, then surround suppression can be thought of as an example of sublinear summation. Similarly, facilitation by the near surround for weak inputs then represents supralinear summation.

We have found that these and other response properties can be understood in some detail from a simple model. We consider a network of excitatory (E) and inhibitory (I) neurons, extended across a 1-D or 2-D space. The strengths of each type of connection – E→E, E→I, I→E, I→I – fall off as functions of distance. As in biology, I projections are shorter-range than E projections. We are guided by previous results showing that the inhibition received by cells is decreased when they are being suppressed by a surround stimulus, relative to their response to a CRF stimulus alone (Ozeki et al. 2009), and correspondingly showing that the firing of inhibitory cells, like that of excitatory cells, is suppressed by surround stimuli (Song and Li 2008). These results led to the conclusion that the E→E connections must be sufficiently strong that, when the network is being driven by a CRF stimulus, they would render the network unstable in the absence of feedback inhibition (Ozeki et al. 2009), a conclusion also supported by other work (London et al. 2010). We

term such a network an inhibition-stabilized network or ISN.

We then add to this the fact that individual neurons have a supralinear, power-law input-output function. This is based on intracellular recordings in anesthetized cat primary visual cortex (V1) showing that a neuron's instantaneous firing rate is well described as a power law function of its instantaneous mean voltage relative to rest (rates and voltages measured in 30 ms bins) with powers ranging from 2 to 5, and that this holds true over the entire dynamic range of neuronal response to visual stimuli (Finn et al. 2007, Priebe and Ferster 2005, 2006, Priebe et al. 2004).[1] This power law relationship is predicted on theoretical grounds when mean input is subthreshold and spiking is driven by input fluctuations (Hansel and van Vreeswijk 2002, Miller and Troyer 2002), as appears to be the case in V1 (Anderson et al. 2000b).

We find that this supralinear network can explain a wide variety of response properties including those described above. Here we mathematically analyze the model. We focus particularly on exposing the origins of the transition in model behavior from supralinear to sublinear summation with increasing input strength, which occurs as the excitatory subnetwork becomes unstable and is stabilized by feedback inhibition. Hence we refer to the network as the stabilized supralinear network or SSN. We also conduct a more detailed analysis of the 2-dimensional case consisting of a single excitatory and a single inhibitory population.

2 Setup: Equations for the Supralinear Network

We take $\mathbf{r} = \begin{pmatrix} \mathbf{r}_E \\ \mathbf{r}_I \end{pmatrix}$ to be the N-dimensional vector of neuronal firing rates, ordered so that the top N_E neurons, represented by $\mathbf{r}_E$, are all excitatory neurons, and the remaining N_I neurons, represented by $\mathbf{r}_I$ are inhibitory neurons, $N_E + N_I = N$. (We refer to the units in our model as "neurons", but, as discussed below, the equations represent average firing rates and so excitatory or inhibitory units may be better understood as local interconnected groups of excitatory or inhibitory neurons, over which the average is taken.) The matrix of connections between the neurons is $\mathbf{W} = \begin{pmatrix} \mathbf{W}_{EE} & -\mathbf{W}_{EI} \\ \mathbf{W}_{IE} & -\mathbf{W}_{II} \end{pmatrix}$ where $\mathbf{W}_{XY}$ is the matrix of connections from neurons of type Y (E or I) to neurons of type X and has non-negative entries. The feedforward input to the neurons in the network is $\mathbf{h} = \begin{pmatrix} \mathbf{h}_E \\ \mathbf{h}_I \end{pmatrix}$.

We study the simplest standard firing-rate-model equations (reviewed in Ermentrout and Terman 2010, Chapter 11; Gerstner and Kistler 2002, Chapter 6; Dayan and Abbott 2001, Chapter

[1] We are assuming that mean voltage is linear in the input. Nonlinearities such as spike-rate-adaptation currents could complicate this picture. We also are ignoring the fact that the power increases with contrast, because the noise level decreases with contrast (Finn et al. 2007), which yields increasing powers (Hansel and van Vreeswijk 2002, Miller and Troyer 2002). However the picture we describe in this paper primarily concerns stabilization against the otherwise explosive nonlinearity of a supralinear input-output function. Thus, the picture should hold so long as the input-output function is supralinear over the cell's dynamic range, as expected for fluctuation-driven spiking – the closer the cell is to threshold, the greater the increase in spiking driven by a given increment of input.

7), in which a neuron's firing rate approaches a nonlinear function of its input with first-order dynamics:

$$\tau \mathbf{T} \frac{d\mathbf{r}}{dt} = -\mathbf{r} + \mathbf{f}(\mathbf{W}\mathbf{r} + \mathbf{h}) \tag{1}$$

Here $\mathbf{T}$ is a diagonal matrix of relative time constants, *i.e.* the time constant of the i^{th} neuron is τT_{ii}. $\mathbf{f}$ is a vector function of a vector argument that acts elementwise on its argument, $(\mathbf{f}(\mathbf{v}))_i = f_i(v_i)$, for some scalar functions of a scalar variable, f_i, where v_i is the i^{th} element of $\mathbf{v}$. These rate model equations do not capture fast time scales that arise in spiking networks, and cannot capture synchronization of spikes across neurons, but tend to be reliable in describing steady states or slower aspects of dynamics when neurons spike asynchronously. We will focus on the steady state and its stability.

We will study the case in which the f_i are identical for all elements, $f_i \equiv f$, and f is a rectified power law with power $n > 1$:

$$f(x) = k([x]_+)^n \tag{2}$$

where $[x]_+ = x$, $x > 0$; $= 0$, otherwise. We will summarize this by saying

$$\tau \mathbf{T} \frac{d\mathbf{r}}{dt} = -\mathbf{r} + k(\mathbf{W}\mathbf{r} + \mathbf{h})^{.n} \tag{3}$$

where $\mathbf{v}^{.n}$ is the vector with i^{th} element $([v_i]_+)^n$ (the period in the exponent $.n$, based on Matlab notation, is to indicate that the operation is done element-by-element rather than to the vector as a whole). A power-law relation between the mean input and mean response arises in the case that spiking is driven by input fluctuations (Hansel and van Vreeswijk 2002, Miller and Troyer 2002), and similarly it is observed in V1 as the relation between the trial-averaged mean voltage and mean response.

We now change variables to dimensionless ones. This allows us to determine the dimensionless combinations of parameters on which model behavior depends and in which expansions for small or large values may be undertaken. We let $\psi = \|\mathbf{W}\|$ where $\|\mathbf{W}\|$ is some matrix norm or other measure of the size of $\mathbf{W}$, and write $\mathbf{W} = \psi\mathbf{J}$ with $\mathbf{J}$ dimensionless and $\|\mathbf{J}\| = 1$. Similarly we let $c = |\mathbf{h}|$ and write $\mathbf{h} = c\mathbf{g}$ with $\mathbf{g}$ dimensionless and $|\mathbf{g}| = 1$ (again, $|\mathbf{g}|$ indicates some measure of the size of a vector, *e.g.* a vector norm). Note that c and $\psi\mathbf{r}$ have the same units, so that $\psi\mathbf{r}/c$ is dimensionless, and that kc^n and $\mathbf{r}$ have the same units, so that $kc^n/(c/\psi) = kc^{n-1}\psi$ is dimensionless. We thus define the dimensionless variable and parameter:

$$\mathbf{y} = \mathbf{r}\psi/c \tag{4}$$

$$\alpha = kc^{n-1}\psi \tag{5}$$

Then equation 3 becomes

$$\tau \mathbf{T} \frac{d\mathbf{y}}{dt} = -\mathbf{y} + \alpha(\mathbf{J}\mathbf{y} + \mathbf{g})^{.n} \tag{6}$$

Thus, given $\mathbf{J}, \mathbf{g}, \mathbf{T}$, and n, the dynamics depends only on the single parameter α.

The fact that Eq. 6 has a single α for all neurons is quite general: if neuron i had parameter α_i, this could be replaced with α by multiplying all weights J_{ij} and inputs g_i to neuron i by $\left(\frac{\alpha_i}{\alpha}\right)^{1/n}$, leaving the form of the equation unchanged. However the fact that the equation has a single n for all neurons is a real restriction. Consideration of n's that vary between neurons or between neuron types remains a question for future study.

Note that we can rewrite the input to $\mathbf{r}$ in our model as $(k^{\frac{1}{n}}\mathbf{W}\mathbf{r} + k^{\frac{1}{n}}c\mathbf{g})^{\cdot n}$, so that the effective recurrent weights are $k^{\frac{1}{n}}\mathbf{W}$ and the effective input strength is $k^{\frac{1}{n}}c$. Then $\alpha = kc^{n-1}\psi = \left(k^{\frac{1}{n}}\|\mathbf{W}\|\right)\left(k^{\frac{1}{n}}c\right)^{n-1}$, that is, α = (recurrent weight)(feedforward weight)$^{n-1}$. Note also that whether the input is dominated by feedforward input $c\mathbf{g}$ or recurrent input $\mathbf{W}\mathbf{r}$ is determined by the size and structure of $\mathbf{y}$ for a given α (because $\mathbf{W}\mathbf{r} + \mathbf{h} = c(\mathbf{J}\mathbf{y} + \mathbf{g})$, so that the balance depends only on the relative sizes of $\mathbf{J}\mathbf{y}$ vs. $\mathbf{g}$), and is not impacted at all by the ratio c/ψ, which naively might be thought to determine the feedforward/recurrent balance. For a given α, this ratio simply scales $\mathbf{r}$ ($\mathbf{r} = (c/\psi)\mathbf{y}$).

We will focus on the equation for the steady-state:

$$\mathbf{y} = \alpha(\mathbf{J}\mathbf{y} + \mathbf{g})^{\cdot n} \tag{7}$$

However, in considering stability of the steady state we will need to use the dynamical equation 6.

3 Scaling Argument

In this section, we show that the supralinear network generically makes a transition from supralinear summation of responses to multiple sets of feedforward inputs for weak inputs ($\alpha \ll 1$) to sublinear summation for stronger inputs ($\alpha \gg 1$), with the transition occuring for α of order of magnitude 1, for which we use the standard notation $\alpha \sim O(1)$. This transition occurs because of dynamic stabilization by feedback inhibition of an otherwise explosive network. We then compare this stabilization to that in the balanced network model of Van Vreeswijk and Sompolinsky (1998).

3.1 Scaling for small α

For $\alpha \ll 1$ we expect the steady state to satisfy $y \approx \alpha\mathbf{g}^{\cdot n}$, since then the $\mathbf{J}\mathbf{y}$ term is small relative to the $\mathbf{g}$ term and so adds only a small correction to this solution. More generally, we can write a formal expression for the steady state:

$$\mathbf{y} = \alpha(\mathbf{g} + \alpha\mathbf{J}(\mathbf{g} + \alpha\mathbf{J}(\mathbf{g} + \alpha\mathbf{J}(\ldots)^{\cdot n})^{\cdot n})^{\cdot n})^{\cdot n} \tag{8}$$

or

$$\mathbf{r} = k(c\mathbf{g} + k\mathbf{W}(c\mathbf{g} + k\mathbf{W}(c\mathbf{g} + k\mathbf{W}(\ldots)^{\cdot n})^{\cdot n})^{\cdot n})^{\cdot n} \tag{9}$$

where the ellipses indicate infinite repetition of the pattern. Assuming quantities in the parentheses are positive so that we can ignore rectification, which they will be for sufficiently small α, equation 8

can be converted into an infinite series in increasing integer powers of α with dominating (lowest-order) term $\alpha \mathbf{g}^{\cdot n}$.[2] The terms multiplying α^p will involve factors of $\mathbf{g}$ interspersed with $\mathbf{J}$'s, with the sum of the powers on the $\mathbf{g}$'s equal to $p(n-1)+1$. Similarly for $\mathbf{r}$, one obtains a series involving an infinite set of powers of c, with lowest-order term $k(c\mathbf{g})^{\cdot n}$ and higher-order terms proportional to $k^p c^{p(n-1)+1}$ and involving a set of $\mathbf{g}$'s with summed power also equal to $p(n-1)+1$. If this series converges, which it will for sufficiently small α, it will give a steady state solution.

Thus, for small α, feedforward inputs sum supralinearly to produce responses. Intuitively: the effective connection between two neurons tells how much the steady-state postsynaptic rate changes for a given change in steady-state presynaptic rate. This is given by the weight between the neurons times the postsynaptic gain. The gain is the slope of the input-output function (Eq. 2), which is monotonically increasing with the postsynaptic cell's firing rate. That is, the steady-state equation is $y_i = \alpha(\sum_j J_{ij} y_j + g_i)^n$, so $\frac{dy_i}{dy_j} = n\alpha(\sum_j J_{ij} y_j + g_i)^{n-1} J_{ij} = n\alpha^{\frac{1}{n}} y_i^{\frac{n-1}{n}} J_{ij}$. For small α, $y_i \approx \alpha g_i^n$ is small, and hence the gain is small. In this regime the network is essentially feedforward driven, with small modifications by the weak effective recurrent connections. Since individual cells respond supralinearly to their inputs, the network sums responses supralinearly.

3.2 Scaling for large α

For sufficiently large α, the series in Eq. 8 will explode rather than converge. We expect this to occur for $\alpha = O(1)$. Physically, this occurs approximately when the effective weights become strong enough that the excitatory subnetwork by itself becomes unstable in the absence of dynamic feedback inhibition. Another way to say this is that inputs are raised to the power $n > 1$ to produce responses which feed back in as inputs; once inputs are sufficiently large, this process is explosive, like a nuclear reaction going critical. If the network nonetheless remains stable, it must be dynamically stabilized by feedback inhibition (Ozeki et al. 2009, Tsodyks et al. 1997).

To be dynamically stabilized, the dependence of $\alpha(\mathbf{J}\mathbf{y}+\mathbf{g})^n$ on the leading α must be cancelled, because otherwise $\mathbf{y} \sim \alpha$, which enters into $\mathbf{J}\mathbf{y}$ and is raised to the n^{th} power to give[3] $\mathbf{y} \sim \alpha^{n+1}$, which enters into $\mathbf{J}\mathbf{y}$, and so on – the infinite series in powers of α results, which will blow up for sufficiently large α. To cancel the leading α, it must be the case that, to leading order in α, $\mathbf{J}\mathbf{y} + \mathbf{g} \sim \alpha^{-\frac{1}{n}}$. This in turn requires that, to leading order, $\mathbf{y}$ has the same α-dependence as $\mathbf{g}$, $\mathbf{y} \sim \alpha^0$, so that the leading order of $\mathbf{y}$ can cancel the $\mathbf{g}$ term leaving only terms of order $\alpha^{-\frac{1}{n}}$. We define

$$\beta = \alpha^{-\frac{1}{n}} \tag{10}$$

Writing $\mathbf{y} = \mathbf{y}_0 + \beta \mathbf{y}_1$, we find the requirement $\mathbf{J}\mathbf{y} + \mathbf{g} \sim \beta$ yields:

$$\mathbf{y}_0 = -\mathbf{J}^{-1}\mathbf{g} \tag{11}$$

$$-\mathbf{J}^{-1}\mathbf{g} + \beta \mathbf{y}_1 = (\mathbf{J}\mathbf{y}_1)^{\cdot n} \tag{12}$$

[2] This can be done by expanding each power in Eq. 8 as $\alpha(\mathbf{g} + \alpha\mathbf{J}(\ldots)^{\cdot n})^{\cdot n} = \alpha(\mathbf{g}^{\cdot n} + n\mathbf{g}^{\cdot(n-1)} .* \alpha\mathbf{J}(\ldots)^{\cdot n} + \frac{n(n-1)}{2}\mathbf{g}^{\cdot(n-2)} .* (\alpha\mathbf{J}(\ldots)^{\cdot n})^{\cdot 2} + \ldots$, where $.*$ indicates element-by-element multiplication of two vectors to create another vector, and then collecting together the terms of each given order in α.

[3] This could be avoided if $\mathbf{J}\mathbf{y} = 0$ to leading order in α, but that requires fine tuning, *i.e.* it requires Det $\mathbf{J} = 0$.

The latter equation shows that $\mathbf{y}_1$ itself has some further dependence on β.

These arguments can be translated in terms of $\mathbf{r}$. Once c is sufficiently large, stability requires cancellation of the linear dependence of $\mathbf{W}\mathbf{r} + c\mathbf{g}$ on c, because otherwise $\mathbf{r} \sim c^n$ which enters back into $\mathbf{W}\mathbf{r}$ and is raised to the n to yield c^{n^2} dependence, and so on. Cancellation requires that, to leading order, $\mathbf{r} \sim c$, which in turn requires that to leading order $\mathbf{W}\mathbf{r} + c\mathbf{g} \sim c^{\frac{1}{n}}$. Writing $\mathbf{r} = c\mathbf{r}_0 + c^{\frac{1}{n}}\mathbf{r}_1$, we find that $\mathbf{r}_0 = \frac{1}{\psi}\mathbf{y}_0 = -\mathbf{W}^{-1}\mathbf{g}$ and $\mathbf{r}_1 = \frac{c}{\psi}\frac{\beta}{c^{\frac{1}{n}}}\mathbf{y}_1 = \frac{1}{\psi(k\psi)^{\frac{1}{n}}}\mathbf{y}_1$, with $\mathbf{r}_1$ satisfying $-\mathbf{W}^{-1}\mathbf{g} + c^{-\frac{n-1}{n}}\mathbf{r}_1 = (\mathbf{W}\mathbf{r}_1)^n$.

These solutions show that, if the network dynamically stabilizes, its responses are a sum of terms that are linear and sublinear in the feedforward inputs, that is, responses can add sublinearly. In studies of 2-dimensional systems (one excitatory and one inhibitory population), we will find that, when the excitatory-neuron element of $-\mathbf{J}^{-1}\mathbf{g}$ is negative, the sublinear term becomes dominant (as it must: $(\mathbf{y}_0)_E < 0$, so one must have $\beta(\mathbf{y}_1)_E > |(\mathbf{y}_0)_E|$ for $y_E > 0$) and network behavior becomes strongly sublinear. In this case, excitatory firing rates eventually peak and then are ultimately pushed to zero with increasing c, *i.e.* with decreasing β, but there is a large dynamic range of c beyond the supralinear-to-sublinear transition before this peak occurs (see Figure 2). The behavior from $c = 0$ until somewhat beyond the peak yields behavior much like that seen in biology, and so we guess that this dynamic range represents the dynamic range of the feedforward input to cortex. This will be discussed in Sections 5.2-5.3.

A more systematic account of the large α (small β) case can be obtained by formulating a solution like Eq. 8 for small α. When the elements of $\mathbf{y}$ are > 0 (more precisely: when the elements of $\mathbf{J}\mathbf{y} + \mathbf{g}$ are ≥ 0), we can rearrange Eq. 7 for the steady state as

$$\mathbf{y} = -\mathbf{J}^{-1}\mathbf{g} + \beta\mathbf{J}^{-1}\mathbf{y}^{\cdot\frac{1}{n}} \tag{13}$$

Then we can formally write a steady-state solution as

$$\mathbf{y} = -\mathbf{J}^{-1}\mathbf{g} + \beta\mathbf{J}^{-1}(-\mathbf{J}^{-1}\mathbf{g} + \beta\mathbf{J}^{-1}(\ldots)^{\cdot\frac{1}{n}})^{\cdot\frac{1}{n}} \tag{14}$$

or

$$\mathbf{r} = -c\mathbf{W}^{-1}\mathbf{g} + \frac{1}{k^{1/n}}\mathbf{W}^{-1}(-c\mathbf{W}^{-1}\mathbf{g} + \frac{1}{k^{1/n}}\mathbf{W}^{-1}(\ldots)^{\cdot\frac{1}{n}})^{\cdot\frac{1}{n}} \tag{15}$$

If quantities in parentheses are positive, a series solution in powers of β can be obtained from Eq. 14 in the same manner as outlined for Eq. 8. When this series converges, which it will for small enough β, it gives a steady-state solution. However, if elements of $\mathbf{J}^{-1}\mathbf{g}$ are negative, then for small enough β the elements in parentheses will no longer be positive (and correspondingly, as mentioned above, in the 2-D case y_E is pushed to zero with decreasing β at finite β, so that Eq. 13 fails at that point). We can instead regard Eq. 13 as an iterative scheme, $\mathbf{y}[p+1] = -\mathbf{J}^{-1}\mathbf{g} + \beta\mathbf{J}^{-1}\mathbf{y}[p]^{\frac{1}{n}}$, beginning from some initial condition $\mathbf{y}[0]$ (Eq. 8 can also be regarded in this way). Writing this as $\mathbf{y}[p+1] = f(\mathbf{y}[p])$, if all of the eigenvalues of the Jacobian of f at the fixed point have absolute values less than 1, then the iteration will converge to the fixed point within some basin of attraction about the fixed point. Hence with suitable initial conditions, one can find solutions through this

iterative scheme, although not for β's less than that at which some elements of $\mathbf{y}$ are pushed to zero.

These scaling arguments provide key insights into the supralinear network (Eq. 3) that is confirmed by other analysis and simulations: for small α, recurrence is weak and the network supralinearly adds responses to different feedforward inputs; with increasing α, there is a transition, for α $O(1)$, to a dynamic stabilization that leads responses to add sublinearly. Note that individual neurons still supralinearly sum the net (feedforward plus recurrent) inputs they receive, but the network "conspires" to deliver net input that is so strongly sublinear that, even after the neuron raises its net input to the power n, its responses add sublinearly. We have found in both high-dimensional and 2-dimensional simulations, and we will show below for the 2-dimensional case, that stabilization will occur provided feedback inhibition is sufficiently strong and the inhibitory time constant is not too slow relative to the excitatory time constant. This transition from supralinear to sublinear behavior in turn appears to underly a wide variety of nonlinearities in neocortical behavior.

3.3 Comparison to the balanced network

Van Vreeswijk and Sompolinsky (1996, 1998) introduced the "balanced network" model (see also Renart et al. 2010). They considered a circuit of stochastic excitatory and inhibitory units that could have firing rate 0 or 1, and studied the conditions in which the network would dynamically find its way to a balanced state in which the mean input is subthreshold, yet firing rates are nonzero, meaning firing is driven by fluctuations. They assumed each unit received K inputs of strength $\frac{1}{\sqrt{K}}$, or a net input of strength $\sqrt{K}$, for K large (*e.g.*, thousands of inputs). The mean field equations for the average E and I firing rates are the 2-dimensional version of the rate equation, Eq. 1, for one E and one I population, where both $\mathbf{W}$ and $c\mathbf{g}$ are of order $\sqrt{K}$,[4] and the function f is a sigmoidal function rising from 0 to 1 as the input moves from approximately -3 to 3, and saturating at 0 or 1 for smaller or larger values respectively. To be in the balanced state, the mean firing rate must be neither 0 nor saturated at 1, so the net input must be $O(1)$ (*i.e.*, between -3 and 3).

Thus, the condition for the balanced state is that $\mathbf{W}\mathbf{r} + c\mathbf{g} \sim O(1)$ where both $\mathbf{W}$ and $c\mathbf{g}$ are $O(\sqrt{K})$. The solution, much as in our scaling argument, is to write $\mathbf{r} = \mathbf{r}_0 + \frac{1}{\sqrt{K}}\mathbf{r}_1 + \ldots$, where $\mathbf{r}_0$ and $\mathbf{r}_1$ are $O(1)$ and the dots represent higher-order terms in $\frac{1}{\sqrt{K}}$. The balanced condition is that the $O(\sqrt{K})$ term in the input vanishes, that is, $\mathbf{W}\mathbf{r}_0 = -c\mathbf{g}$, leaving as input only the $O(1)$ term $\frac{\mathbf{W}}{\sqrt{K}}\mathbf{r}_1$ and terms that are $O(\frac{1}{\sqrt{K}})$.

The condition for dynamic stabilization to achieve the balanced state, $\mathbf{r}_0 = -c\mathbf{W}^{-1}\mathbf{g}$, is of course identical to the condition we have found for dynamic stabilization in our model.[5] Although the condition is formally identical, the meaning is different in crucial ways:

[4] The input is expressed in units of a variance term which itself is dynamically determined, but this term is $O(1)$ and does not impact the points made here.

[5] We wrote the leading term as $c\mathbf{r}_0$ rather than $\mathbf{r}_0$, but the leading terms are identical.

1. In the balanced model, the stabilization is required because inputs are large and must cancel to leave something small. In our model, the stabilization arises because the dynamics are explosive, due to the supralinear input-output function, and occurs when none of the inputs are large. This can be seen by recalling that in our model the effective recurrent weights are $k^{\frac{1}{n}}\mathbf{W}$ and the effective input strength is $k^{\frac{1}{n}}c$, and that $\alpha = \left(k^{\frac{1}{n}}\|\mathbf{W}\|\right)\left(k^{\frac{1}{n}}c\right)^{n-1}$. In the balanced network, the recurrent weights $k^{\frac{1}{n}}\|\mathbf{W}\|$ and the feedforward weight $k^{\frac{1}{n}}c$ are both $O(\sqrt{K})$, so α is $O(K^{\frac{n}{2}})$. In contrast, stabilization arises in our model when α is $O(1)$.

2. In the balanced network, the second-order term $\frac{1}{\sqrt{K}}\mathbf{r}_1$ is negligibly small relative to the first-order term $\mathbf{r}_0$ (because the stabilization is to cancel large things, leaving something small). The first-order term is linear in the input, $\mathbf{r}_0 = -c\mathbf{W}^{-1}\mathbf{g}$, and so in the balanced network responses are always linear in the input. In our model, because the stabilization occurs when inputs are small, the second-order term can be comparable to or larger than the first-order term over a wide dynamic range, enabling a variety of sublinear behavior.

   In particular, in our model elements of $\mathbf{r}_0$ can be negative, meaning that for such an element $r_1 > |r_0|$ over the relevant dynamic range of behavior (discussed in more detail for the 2-D model in Section 5.2). In the balanced model, since all terms except $\mathbf{r}_0$ are negligible, the elements of $\mathbf{r}_0$ must be positive for activity to be nonzero.

In sum, in the balanced network, inputs are huge relative to the distance from rest to threshold, and must dynamically cancel for the network to neither saturate nor have 0 activity but instead be in the fluctuation-driven regime. The dynamic cancellation or stabilization yields network responses that are always linear in the input. In our model, the supralinear input-output function renders the network explosive – input is raised to a power greater than 1 to produce responses, which feed back as input. Stabilization against this explosive nonlinearity arises when inputs are relatively small, yielding a range of sublinear behavior.

4 Reduction to a 2-dimensional system

Most of our analysis hereafter will focus on a 2-dimensional system of one excitatory and one inhibitory population, as it is difficult to say much in general in higher dimensions. A 2-D system of one E and one I population can be derived as a mean field equation from higher-dimensional models in which E and I neurons have random connectivity (*e.g.* Renart et al. 2010, van Vreeswijk and Sompolinsky 1998). In particular, if the high-dimensional model involves integrate-and-fire neurons, their input-output functions in the fluctuation-driven regime can be reasonably approximated by power-law functions (Hansel and van Vreeswijk 2002).

Here we consider a higher-dimensional system with structured connectivity. We show a heuristic derivation of a 2-D system that preserves a surprising amount of the behavior of the higher-dimensional system. We then show how the conditions for "normalization" – sublinear addition of

responses to multiple stimuli – in the high-D system can be expressed as simple conditions in the 2-D system on the growth of $\mathbf{r}$ with increasing c, or the growth of $\mathbf{y}$ with increasing α.

4.1 Reduction

We consider a topographic network, with pairs of excitatory (E) and inhibitory (I) units arranged on a 1-D or 2-D grid with periodic boundary conditions. Stimuli are localized on the grid, though there may be more than one localized stimulus present. For simplicity, we assume a single time constant for all E cells and one for all I cells. We let θ represent the position on the grid, and $r_E(\theta)$ and $r_I(\theta)$ the excitatory and inhibitory firing rates at position θ. Thus, we can write Eq. 3 as

$$\tau_E \frac{dr_E(\theta)}{dt} = -r_E(\theta) + k\left(\psi J_{EE} * r_E(\theta) - \psi J_{EI} * r_I(\theta) + cg_E(\theta)\right)_{\cdot}^{n} \quad (16)$$

$$\tau_I \frac{dr_I(\theta)}{dt} = -r_I(\theta) + k\left(\psi J_{IE} * r_E(\theta) - \psi J_{II} * r_I(\theta) + cg_I(\theta)\right)_{\cdot}^{n} \quad (17)$$

Here, $J_{XY} * r_Y(\theta) = \sum_{\theta'} J_{XY}(\theta, \theta') r_Y(\theta')$.

We will consider "normalization", the sublinear addition of the responses to two stimuli (Carandini and Heeger 2011). We let one stimulus be centered at $\theta = 0$. We let $\vec{j}_{XY}$ be the vector of weights $J_{XY}(0, \theta)$ to position 0, normalized so that the weight from position 0 is 1, and let $\tilde{J}_{XY} = J_{XY}(0, 0)$, the actual value of the weight from position 0. Similarly, we let $\hat{\mathbf{r}}_E$, $\hat{\mathbf{r}}_I$ be the vectors of excitatory and inhibitory firing rates, respectively, normalized to equal 1 at position 0. Then the equations for the units at position 0 are

$$\tau_E \frac{dr_E(0)}{dt} = -r_E(0) + k\left(\psi \tilde{J}_{EE} r_E(0)\left(\vec{j}_{EE} \cdot \hat{\mathbf{r}}_E\right) - \psi \tilde{J}_{EI} r_I(0)\left(\vec{j}_{EI} \cdot \hat{\mathbf{r}}_I\right) + cg_E(0)\right)_{\cdot}^{n} \quad (18)$$

$$\tau_I \frac{dr_I(0)}{dt} = -r_I(0) + k\left(\psi \tilde{J}_{IE} r_E(0)\left(\vec{j}_{IE} \cdot \hat{\mathbf{r}}_E\right) - \psi \tilde{J}_{II} r_I(0)\left(\vec{j}_{II} \cdot \hat{\mathbf{r}}_I\right) + cg_I(0)\right)_{\cdot}^{n} \quad (19)$$

Although we had previously incorporated changes in $|\mathbf{g}|$ into c, we now take addition of a second stimulus to simply alter $\mathbf{g}$, so that in particular addition of a second stimulus that gives no input to position 0 does not alter $g_E(0)$ or $g_I(0)$.

We now make the ansatz that, as the stimulus changes, the four dot products are all scaled by a common factor, whether the stimulus changes in strength (changing c) or in shape (changing $\mathbf{g}$, *e.g.* adding a second stimulus). We can then write $J_{XY} = \tilde{J}_{XY} \frac{\left(\vec{j}_{XY} \cdot \hat{\mathbf{r}}_Y\right)}{\left(\vec{j}_{EE} \cdot \hat{\mathbf{r}}_E\right)}$ and $\Psi \equiv \psi\left(\vec{j}_{EE} \cdot \hat{\mathbf{r}}_E\right)$, where Ψ includes the common scaling factor (note that the matrix $\mathbf{J}$ of these J's need not satisfy $\|\mathbf{J}\| = 1$). Since the weights $\vec{j}_{XY}$ are non-negative, the effect of adding a second non-negative stimulus is to increase Ψ, hence "normalization" of the E or I population corresponds to a decrease in firing rates $r_E(0)$ or $r_I(0)$, respectively, with increasing Ψ: $\frac{dr_E(0)}{d\Psi} < 0$, $\frac{dr_I(0)}{d\Psi} < 0$.

With this ansatz, we obtain 2-dimensional equations. Letting $r_E \equiv r_E(0)$, $g_E \equiv g_E(0)$, etc., we

obtain the equations

$$\tau_E \frac{dr_E}{dt} = -r_E + k\left(\Psi J_{EE} r_E - \Psi J_{EI} r_I + c g_E\right)^{.n} \quad (20)$$

$$\tau_I \frac{dr_I}{dt} = -r_I + k\left(\Psi J_{IE} r_E - \Psi J_{II} r_I + c g_I\right)^{.n} \quad (21)$$

This is just the 2-dimensional version of Eq. 3, and is equivalent to Eq. 6 for 2-dimensional $\mathbf{y}$ with Ψ replacing ψ in the definitions of $\mathbf{y}$ and α (Eq. 5).

The ansatz, of course, is not in general true, but it can be close enough to true to give a good account of the higher-dimensional system. To illustrate this, we simulate the model on a one-dimensional ring, which we can think of as representing preferred orientation.[6] We consider 180 E/I pairs at grid positions separated by 1° in preferred orientation, with 0° = 180°. All four connection types have the same width, following evidence that excitatory and inhibitory inputs received by cells in upper layers have similar orientation tuning (Anderson et al. 2000a, Ferster 1986, Marino et al. 2005, Martinez et al. 2002). The connectivity takes the form

$$W_{XY}(\theta, \theta') = J_{XY} e^{-\frac{d(\theta,\theta')^2}{2\sigma_{\text{ori}}^2}} \quad (22)$$

where $d(\theta, \theta')$ is the shortest distance around the circle between θ and θ', $J_{EE} = 0.0441$, $J_{IE} = 0.04158$, $J_{EI} = 0.0231$, and $J_{II} = 0.01827$. (Here we are not sticking to our convention of $\|\mathbf{J}\| = 1$, or in other words, we are taking $\psi = 1$ with these values for $\mathbf{J}$.) We take $\sigma_{\text{ori}} = 32°$, but also consider other values in Fig. 1C. Other parameters are $\tau_E = 20$ ms, $\tau_I = 10$ ms, $k = 0.04$, and $n = 2.0$. We consider stimulation by either one oriented luminance grating or two orthogonal gratings. Each grating is represented by a Gaussian-shaped curve of feedforward input with width (standard deviation of the Gaussian) 30° and height c; a single grating is centered at $\theta = 0°$, a second added grating is centered at $\theta = 90°$. We also consider varying the width of the feedforward Gaussian, for a grating centered at $\theta = 0°$. For any given stimulus (1 or 2 stimuli, stimulus height c, given stimulus width) the equivalent 2-D model is found as follows: we use the same J's and other parameters, and take Ψ to be the value of the convolution, at $\theta = 0$, of the connectivity Gaussian (Eq. 22 with $J_{XY} = 1$) with the input pattern for $c = 1$, which we use as a surrogate for the shape of the response. For the default Gaussian width, this yields $\Psi = 54.86$ for one stimulus and $\Psi = 67.67$ for two stimuli.

The result is that the reduced 2-D model accurately reproduces the behavior of the full model (Fig. 1). The firing rates of the cells at $\theta = 0$ vs. stimulus strength closely match the firing rates in the 2-D model (Fig. 1A). Both models show a similar transition from supralinear summation of responses to the two gratings for weak stimuli to sublinear summation or "normalization" for

[6]The paradigm we study here – suppression of response to one orientation by presentation of an orthogonal orientation – is known as "cross-orientation suppression". In V1, this appears to be primarily mediated by sublinear addition of the feedforward inputs to V1 evoked by the two stimuli (Lauritzen et al. 2001, Li et al. 2006, Priebe and Ferster 2006). However we use this paradigm to study how the model cortex sums responses to multiple stimuli, assuming the feedforward inputs sum linearly.

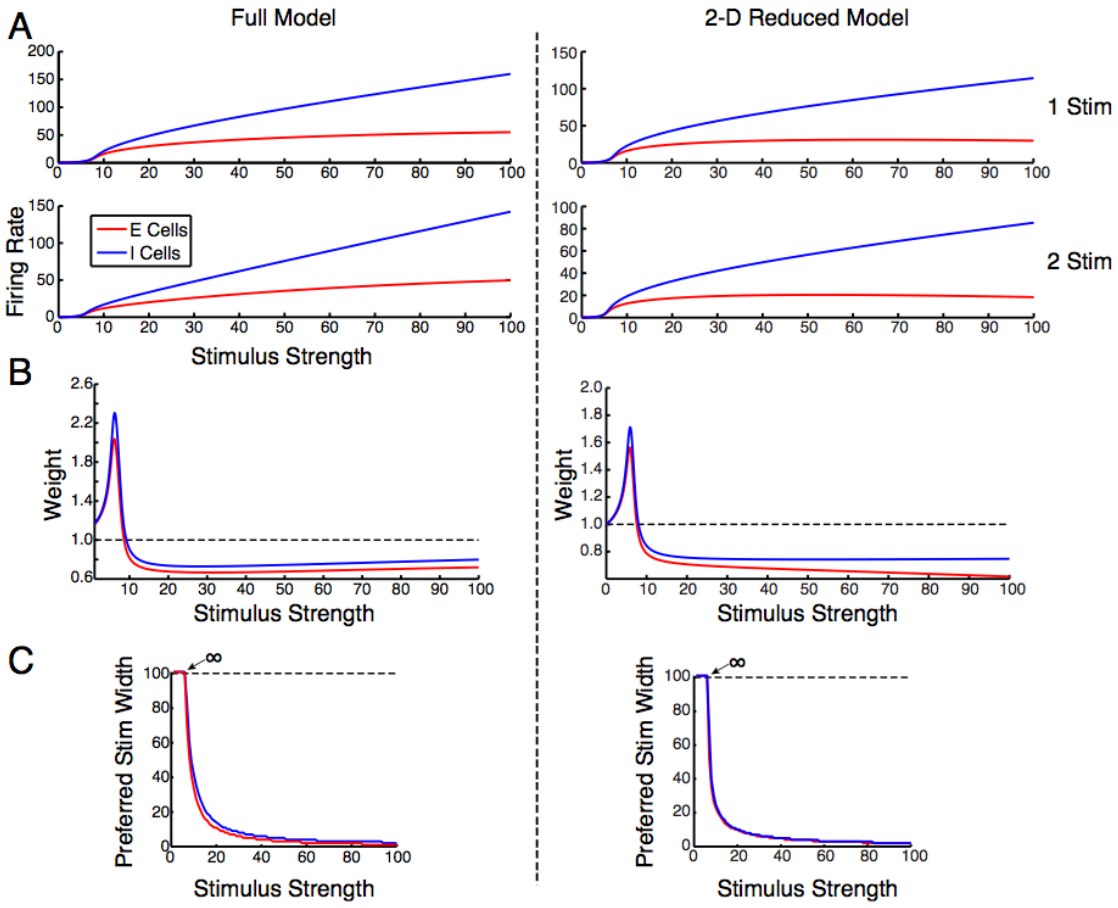

Figure 1:

Two neuron approximation (right) of the full ring model (left). (A) The reduced version of the model produces qualitatively similar curves of response vs. stimulus strength c (for the full model, this is the response of the cells at $\theta = 0$). (B) Full and reduced models show a similar stimulus-strength-dependent transition from supralinear summation (weight > 1) to sublinear summation (weight < 1) of the responses to two gratings. For the full model, for either E or I cells, we let $R_1(\theta)$, $R_2(\theta)$, and $R_{12}(\theta)$ be the response to one grating, the other grating, or the superposition of the two, and find the weight w that gives the least-squares-fit to $R_{12}(\theta) = w\,(R_1(\theta) + R_2(\theta))$. For the reduced model, the corresponding responses are R_1, R_2, and R_{12}, and the weight is $w = \frac{R_{12}}{R_1+R_2}$. (C) Full and reduced models have nearly identical stimulus-strength-dependent tuning for the width of a feedforward stimulus (full model: width of Gaussian stimulus centered at $\theta = 0$ with given stimulus strength c that gives strongest response in cells at $\theta = 0$; reduced model: Ψ is computed for each stimulus width, and plot shows width whose Ψ gives maximal response). In all curves, red shows E cells and blue shows I cells. All responses are steady-state responses. Full model solutions found by simulating until convergence to steady state.

stronger stimuli (Fig. 1B). The network also shows a form of surround suppression, in which the "summation field size" – the stimulus width that yields maximal response for a given stimulus strength – shrinks monotonically with increasing stimulus strength, as is well known in real space (rather than orientation space) for V1 cells (Cavanaugh et al. 2002, Sceniak et al. 1999), and this behavior is extremely similar in the full and reduced models (Fig. 1C).[7] Thus, the 2D model can provide a good basis for understanding more general models.

4.2 Conditions for Normalization

We consider steady-state $\mathbf{r}$ or $\mathbf{y}$ and use expressions like $\frac{dr_X}{d\psi}$ to refer to the dependence of the steady state on parameters. We have seen that the condition that r_X ($X \in \{E, I\}$) exhibits normalization in the high-D model is equivalent to the condition $\frac{dr_X}{d\Psi} < 0$ in the 2-D model. Since Eqs. 20-21 are equivalent to Eqs. 5-6 with Ψ replacing ψ, we revert to the notation of Eqs. 3-6 and use ψ.

We work with the 2-D model and express the conditions $\frac{dr_X}{d\psi} < 0$ as a single vector condition. We note first that $\frac{d\mathbf{r}}{d\psi} = c\frac{d\mathbf{y}/\psi}{d\psi} = c\left(\frac{1}{\psi}\frac{d\mathbf{y}}{d\psi} - \frac{\mathbf{y}}{\psi^2}\right)$ and $\frac{d\mathbf{y}}{d\psi} = \frac{d\mathbf{y}}{d\alpha}\frac{d\alpha}{d\psi} = kc^{n-1}\frac{d\mathbf{y}}{d\alpha}$. Putting these together we find $\frac{d\mathbf{r}}{d\psi} = \frac{kc^n}{\psi}\left(\frac{d\mathbf{y}}{d\alpha} - \frac{\mathbf{y}}{\alpha}\right)$. Thus, the condition for normalization is that $\mathbf{y}$ grow more slowly than linearly with increasing α: $\frac{d\mathbf{y}}{d\alpha} < \frac{\mathbf{y}}{\alpha}$ or $\frac{d\ln\mathbf{y}}{d\ln\alpha} < 1$ or, roughly, that $\mathbf{y} \sim \alpha^p$ for $p < 1$. As we have seen, p becomes less than 1 precisely when the transition from the supralinear to the sublinear scaling regime occurs.

We can reexpress this in terms of $\mathbf{r}$. Using algebra similar to the above, we find $\frac{d\mathbf{r}}{dc} = \frac{(n-1)\alpha}{\psi}\left(\frac{d\mathbf{y}}{d\alpha} + \frac{\mathbf{y}}{(n-1)\alpha}\right)$, from which we find that $\frac{d\mathbf{y}}{d\alpha} < \frac{\mathbf{y}}{\alpha}$ is equivalent to $\frac{d\mathbf{r}}{dc} < n\frac{\mathbf{r}}{c}$. Thus, the condition for normalization is that $\mathbf{r}$ grow more slowly than c^n with increasing c: $\frac{d\ln\mathbf{r}}{d\ln c} < n$ or,

[7] Note that this "summation field size" for orientation selectivity should not be confused with the orientation tuning width, which is the width of the orientation tuning curve obtained by studying response vs. single orientations (more precisely: studying response vs. center orientation, using stimuli that evoke a fixed curve of feedforward input vs. orientation that is symmetric about the center orientation). The orientation tuning curve, representing the set of single orientations that can drive the cell, is analogous to the "minimal response field" in real space, which represents the sum of the set of small regions in visual space in which appropriate light stimuli can evoke spiking responses. The minimal response field in real space is invariant with stimulus contrast (Song and Li 2008), and so too is the shape of the orientation tuning curve (Anderson et al. 2000b, Ferster and Miller 2000, Skottun et al. 1987) (contrast is monotonically related to the firing rate of the inputs to cortex (*e.g.* Ohzawa et al. 1985)). The fact that the summation field size in real space is larger than the minimal response field indicates that stimuli in regions where light cannot directly drive spikes can facilitate responses to stimuli in the minimal response field. Recall that the size of this facilitating area shrinks with contrast. The model suggests that the same may be true in the orientation domain, in terms of cortical processing of feedforward input to cells of different preferred orientations. However, attempts to test this idea will likely be compromised by two facts: (1) simultaneous presentation of multiple orientations does not yield linear summation of the input to cortex evoked by the individual orientations (Lauritzen et al. 2001, Li et al. 2006, Priebe and Ferster 2006) and (2) varying the feedforward orientation tuning by changing stimulus attributes – *e.g.* a sinusoidal luminance grating of a given size provides drive to cortical cells with an orientation tuning that narrows with increasing spatial frequency, and similarly a longer bar drives narrower orientation tuning than a shorter bar – also changes other attributes to which the neurons are independently sensitive, such as spatial frequency or bar length.

roughly, that $\mathbf{r} \sim c^p$ for $p < n$. Again, p becomes less than n at the transition from supralinear to sublinear scaling.

Finally, noting that the steady state condition is $\mathbf{r} = k(\mathbf{J}\mathbf{r} + c\mathbf{g})^{\cdot n}$, without loss of generality we write $\mathbf{J}\mathbf{r} = c\mathbf{f}(c)$ for some vector function $\mathbf{f}$ of c, so that the steady state condition $\mathbf{r} = k(\mathbf{J}\mathbf{r} + c\mathbf{g})^{\cdot n}$ becomes $\mathbf{r} = c^n(\mathbf{f}(c) + \mathbf{g})^{\cdot n}$. Thus we see that a component of $\mathbf{r}$ grows more slowly than c^n precisely when the corresponding component of $\mathbf{f}(c)$ is a decreasing function of c (that is, for corresponding components r and f, $\frac{dr}{dc} < n\frac{r}{c}$ precisely when $f'(c) < 0$). Thus, the condition for normalization can alternatively be expressed as the requirement that $\mathbf{J}\mathbf{r}$ grow more slowly than linearly with c, *i.e.* $\frac{d\mathbf{J}\mathbf{r}}{dc} < \frac{\mathbf{J}\mathbf{r}}{c}$ or $\frac{d\ln(\mathbf{J}\mathbf{r})}{d\ln c} < 1$.

5 Analyses of the 2-Dimensional Network

We will assume throughout this analysis that $g_E \geq 0$, $g_I \geq 0$. We will use the following definitions:

$$\Omega_E \equiv \operatorname{Det}\mathbf{J}\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E = J_{II}g_E - J_{EI}g_I \tag{23}$$

$$\Omega_I \equiv \operatorname{Det}\mathbf{J}\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I = J_{IE}g_E - J_{EE}g_I \tag{24}$$

We also note that there are three possible conditions: (1) $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E > 0$ and $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I > 0$; (2) $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E < 0$ and $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I > 0$; and (3) $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E < 0$ and $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I < 0$. The 4^{th} condition, $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E > 0$ and $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I < 0$, is not mathematically possible for $g_E \geq 0$ and $g_I \geq 0$: $\Omega_E > 0$ and $\Omega_I < 0$ together imply $\operatorname{Det}\mathbf{J} < 0$, and similarly $\Omega_E < 0$ and $\Omega_I > 0$ together imply $\operatorname{Det}\mathbf{J} > 0$.

5.1 When Does the Network Dynamically Stabilize?

5.1.1 The case of infinitely fast inhibition

We first analyze the case of infinitely fast inhibition, $\tau_I/\tau_E = 0$, with constant feedforward inputs. We show that in this case the network, if $\operatorname{Det}\mathbf{J} > 0$, the network is always driven to a stable fixed point from arbitrary starting conditions. The condition $\operatorname{Det}\mathbf{J} > 0$ means that feedback inhibition is sufficiently strong: $J_{EI}J_{IE} > J_{EE}J_{II}$. In addition, we show that if $\operatorname{Det}\mathbf{J} < 0$, sufficiently large initial firing rates will cause the system to "blow up", *i.e.* firing rates will grow arbitrarily large.

With $\tau_I = 0$, the value of y_I is "slaved" to, or instantaneously set by the value of, y_E according to the $\frac{dy_I}{dt}$ part of Eq. 6 for $\mathbf{y}$. Because of the nonlinearity, we cannot solve this for y_I as a function of y_E, but we can instead solve for y_E as a function of y_I:

$$y_E = \frac{1}{J_{IE}}\left(\left(\frac{y_I}{\alpha}\right)^{\frac{1}{n}} + J_{II}y_I - g_I\right) \tag{25}$$

Substituting this in the $\frac{dy_E}{dt}$ part of Eq. 6 yields, after a bit of algebra, an equation for $\frac{dy_I}{dt}$ induced

by the slaving of y_I to the y_E dynamics:

$$\tau_E \frac{dy_I}{dt} = \frac{n\alpha^{\frac{1}{n}} y_I^{\frac{n-1}{n}}}{1 + J_{II} n\alpha^{\frac{1}{n}} y_I^{\frac{n-1}{n}}} \left(-J_{II} y_I - \left(\frac{y_I}{\alpha}\right)^{\frac{1}{n}} + g_I + \frac{\alpha}{J_{IE}^{n-1}} \left(-\mathrm{Det}\,\mathbf{J}\, y_I + J_{EE} \left(\frac{y_I}{\alpha}\right)^{\frac{1}{n}} + \Omega_E \right)^{\cdot n} \right) \tag{26}$$

For sufficiently large y_I, if $\mathrm{Det}\,\mathbf{J} > 0$, the term inside the parentheses in the $\frac{\alpha}{J_{IE}^{n-1}} (\ldots)^{\cdot n}$ term will be negative, and so will be set to zero after the thresholding involved in the $()^{\cdot n}$ operation. The dominant term will then be the $-J_{II} y_I$ term, which is negative. So for sufficiently large y_I, $\frac{dy_I}{dt} < 0$. On the other hand, if $\mathrm{Det}\,\mathbf{J} > 0$, then for sufficiently large y_I, the $(\ldots)^{\cdot n}$ will be positive and larger than the sum of the other terms, so that $\frac{dy_I}{dt} > 0$ and, since increasing y_I will increase $\frac{dy_I}{dt} > 0$, this derivative is ever-increasing.

For sufficiently small y_I, $\frac{dy_I}{dt} > 0$ if either g_I or g_E is nonzero, which can be seen as follows. For sufficiently small y_I, the source terms g_I and Ω_E, if nonzero, dominate the terms involving y_I. Both g_I and the $(\ldots)^{\cdot n}$ term containing Ω_E are non-negative, so if either is positive $\frac{dy_I}{dt}$ will be positive; if $\Omega_E > 0$, the $(\ldots)^{\cdot n}$ term is positive; if $\Omega_E \leq 0$, this implies $g_I > 0$ (given that at least one of g_I and g_E is nonzero, and that both are non-negative).

Thus, for $\mathrm{Det}\,\mathbf{J} > 0$, y_I is driven to a stable fixed point, and y_E is then determined from Eq. 25, so the system will arrive at a stable fixed point. Note that the system could have multiple fixed points with varying levels of y_I. The topology of flow along the y_I axis tells us that there must be an odd number of fixed points, alternating from stable to unstable to stable with increasing y_I, with the outermost fixed points (those with lowest and highest y_I) being stable. In the simplest case, there is a single stable fixed point. In addition, for $\mathrm{Det}\,\mathbf{J} < 0$, the system will blow up for sufficiently large initial firing rates.

5.1.2 More general requirements for stability

Changes in the time constants can alter the stability of the fixed points, but do not alter the number or positions of the fixed points. The results of the previous section tells us that, for $\mathrm{Det}\,\mathbf{J} > 0$, the system always has a fixed point that is stable for $\tau_I = 0$. We consider such a fixed point, and ask when it retains or loses stability for finite τ_I.

We let the fixed point be $\begin{pmatrix} y_E \\ y_I \end{pmatrix}$, and assess stability by linearizing the dynamics about this fixed point. We let $q = \tau_I/\tau_E > 0$. Setting $\tau = \tau_E$ in Eq. 6, the matrix $\mathbf{T}$ is given by $\mathbf{T} = \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix}$. Define the matrix $\mathbf{\Phi} = n\alpha^{\frac{1}{n}} \begin{pmatrix} y_E^{\frac{n-1}{n}} & 0 \\ 0 & y_I^{\frac{n-1}{n}} \end{pmatrix}$. Writing the identity matrix as $\mathbf{1}$, the Jacobian matrix of the 2-D system is:

$$\mathcal{J} = \mathbf{T}^{-1} (\mathbf{\Phi J} - \mathbf{1}) \tag{27}$$

A fixed-point of the dynamics will be stable if $\mathcal{J}$ has a negative trace and a positive determinant.

The negative trace condition is $\mathcal{J}_{EE} < \mathcal{J}_{II}$, which becomes

$$\mathcal{J}_{EE} = n\alpha^{\frac{1}{n}} y_E^{\frac{n-1}{n}} J_{EE} - 1 \leq 0 \text{ OR } \left(\mathcal{J}_{EE} > 0 \text{ AND } q < \frac{n\alpha^{\frac{1}{n}} y_I^{\frac{n-1}{n}} J_{II} + 1}{n\alpha^{\frac{1}{n}} y_E^{\frac{n-1}{n}} J_{EE} - 1} \right) \tag{28}$$

The condition $\mathcal{J}_{EE} \leq 0$ means that the excitatory subnetwork by itself is stable (or marginally stable), which guarantees that the network will always be stable, since only excitatory instability can destabilize the network. When the excitatory subnetwork is unstable, we can further reduce the condition on q for $n = 2$:[8]

$$q < \frac{\sqrt{1 + 4\alpha J_{II}\left(g_I + J_{IE} y_E\right)}}{2J_{EE}\sqrt{\alpha y_E} - 1} \qquad (n = 2,\ \ 2\sqrt{\alpha y_E} J_{EE} > 1) \tag{29}$$

The determinant condition, $\operatorname{Det} \mathcal{J} > 0$, is always true for any fixed point that is stable at $q = 0$. To see this, note that the sign of the determinant does not depend on q for $q > 0$ (because $\operatorname{Det} \mathbf{AB} = \operatorname{Det} \mathbf{A} \operatorname{Det} \mathbf{B}$ for any matrices $\mathbf{A}, \mathbf{B}$, and $\operatorname{Det} \mathbf{T}^{-1} = \frac{1}{q}$). So if we prove that $\operatorname{Det} \mathcal{J} > 0$ for arbitrarily small $q > 0$, we will have shown that it holds for all $q > 0$. For $q = 0$, the determinant, which is the product of the two eigenvalues, was infinite: because the fixed point was stable, both eigenvalues had negative real part: one real part was infinite, corresponding to the instantaneous flow onto the inhibitory nullcline (the line in the y_E/y_I plane on which $\frac{dy_I}{dt} = 0$); the other was finite, corresponding to the flow along the nullcline converging onto the fixed point. (Since the two real parts were unequal, both eigenvalues were real.) As q is moved infinitesimally from 0, the infinite eigenvalue becomes a large but finite negative eigenvalue, while the finite eigenvalue is perturbed by arbitrarily small amounts as q is made arbitrarily small. This means that there is a range of $q > 0$ for which the eigenvalues continue to have negative real parts, and therefore for which the determinant condition holds. Therefore, the determinant condition holds for all q. Thus, for a fixed point that is stable for $q = 0$, the fixed point remains stable so long as condition 28, or condition 29 for $n = 2$, is satisfied.

We also note that, for the case $n = 2$ and for $q \leq 1$, a sufficient condition to conclude that there is only a single fixed point, which is stable, is $\operatorname{Det} \mathbf{J} > 0$ and $J_{EE}^2 < J_{IE} J_{II}$, which can be seen as follows. The determinant condition is $\operatorname{Det}\left(\mathbf{\Phi J} - \mathbf{1}\right) > 0$. We note that, for an arbitrary 2-dimensional matrix $\mathbf{M}$, $\operatorname{Det}\left(\mathbf{M} - \mathbf{1}\right) = \operatorname{Det} \mathbf{M} - \operatorname{Tr} \mathbf{M} + 1$. Thus, the determinant condition is $\operatorname{Det} \mathbf{\Phi J} > \operatorname{Tr} \mathbf{\Phi J} - 1$. Since $\operatorname{Det} \mathbf{\Phi} > 0$ (because firing rates and α are > 0), this condition will be satisfied if $\operatorname{Det} \mathbf{J} > 0$ and $\operatorname{Tr} \mathbf{\Phi J} < 1$. The trace condition for stability is $\operatorname{Tr} \mathbf{T}^{-1} \mathbf{\Phi J} < 1 + q$. But, for $q \leq 1$ and given the structures of $\mathbf{\Phi}$ and $\mathbf{J}$, $\operatorname{Tr} \mathbf{T}^{-1} \mathbf{\Phi J} \leq \operatorname{Tr} \mathbf{\Phi J}$, so the condition $\operatorname{Tr} \mathbf{\Phi J} < 1$ ensures that the trace condition is also satisfied. This condition is

$$J_{EE} y_E^{\frac{n-1}{n}} - J_{II} y_I^{\frac{n-1}{n}} < \frac{1}{n\alpha^{\frac{1}{n}}} \tag{30}$$

[8] This condition is found by solving $\sqrt{y_I} = \sqrt{\alpha}\left(J_{IE} y_E - J_{II} y_I + g_I\right)$ as a quadratic equation for $\sqrt{y_I}$. Discarding the negative solution, this yields $\sqrt{y_I} = \frac{-1 + \sqrt{1 + 4\alpha J_{II}\left(g_I + J_{IE} y_E\right)}}{2J_{II}\sqrt{\alpha}}$. Substituting this into Eq. 28 for $n = 2$ yields Eq. 29.

For $n = 2$, we substitute the solution for $\sqrt{y_I}$ as a function of y_E (footnote 8) into Eq. 30 for $n = 2$ to find $J_{EE}^2 - J_{IE}J_{II} < \frac{1+4\alpha g_I J_{II}}{4\alpha y_E}$. Since the right side is positive, a sufficient condition for this to be true is $J_{EE}^2 < J_{IE}J_{II}$. Recall that, if there is more than one fixed point, some will be unstable at $q = 0$, and they must remain unstable for some region of small but finite q. Since this condition guarantees that any fixed point is stable, we conclude that there can only be one fixed point, which is stable, when this condition holds.

In summary, for $q = \tau_I/\tau_E = 0$, the network always flows to a stable fixed point if $\mathrm{Det}\,\mathbf{J} > 0$. For $q > 0$, a fixed point that is stable at $q = 0$ remains stable when Eq. 28 or, for $n = 2$, Eq. 29 is satisfied. Note that this condition does not ensure that the network always flows to a stable fixed point; for nonzero q there may be initial conditions outside the basin of attraction of the stable fixed point or points. A condition that ensures that any fixed point is stable for $q \leq 1$, and therefore that there is only one fixed point, is $\mathrm{Det}\,\mathbf{J} > 0$ and $J_{EE}^2 < J_{IE}J_{II}$. If this fixed point is the only attractor (there are no limit cycles), this ensures that the network will flow to the stable fixed point. Excepting Eqs. 28-29, these conditions involve feedback inhibition being sufficiently strong: $J_{IE}J_{EI} > J_{EE}J_{II}$, and $J_{IE} > \frac{J_{EE}^2}{J_{II}}$.

In Fig. 2, bottom row, we will illustrate the range of q's yielding stability for various parameter choices with $n = 2$.

5.2 The case $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E < 0$ and supersaturation

We consider Eq. 3 for $\mathbf{r}$, but substituting $\psi\mathbf{J}$ for $\mathbf{W}$. We restrict to the case $\mathrm{Det}\,\mathbf{J} > 0$, which ensures a stable fixed point for at least some range of $\frac{\tau_I}{\tau_E} > 0$. We note that for $\mathrm{Det}\,\mathbf{J} > 0$, $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E < 0$ and $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_I < 0$ are equivalent to $\Omega_E < 0$ and $\Omega_I < 0$, respectively.

We shall equate increasing or decreasing c with increasing or decreasing stimulus contrast. This is based on the fact that the contrast of a visual stimulus is monotonically (but nonlinearly) related to the firing rate of the inputs to V1 from the lateral geniculate nucleus (LGN) (*e.g.* Ohzawa et al. 1985).

In simulations, we find that if $\Omega_E < \Omega_I < 0$ for $g_E = g_I$, then r_E grows with c for a range of c considerably beyond the transition from supralinear to sublinear behavior, but ultimately peaks and is pushed back to 0 with increasing c (see Fig. 2A). The inputs to cortex have limited dynamic range (*e.g.*, stimulus contrast cannot increase beyond 100%), and so we imagine that this circuit may model cortex but that the maximal input strength seen biologically cannot drive excitatory responses too far beyond their peak. The decrease in response with increasing contrast after a peak response is referred to as "supersaturation", and is seen in virtually all V1 cells for contrasts larger than about 75% (Ledgeway et al. 2005, Li and Creutzfeldt 1984, Peirce 2007, Tyler and Apkarian 1985). This model behavior provides one possible explanation for supersaturation, although supersaturation might also in part reflect a supersaturation of inputs, *e.g.* if feedforward inhibition (Bruno 2011) overtakes feedforward excitation with increasing contrast.

Here we analyze this behavior. We shall find that (1) if $\mathbf{r}$ is a stable fixed point, then $\frac{dr_E}{dc}$ and $\frac{dr_I}{dc}$ are negative precisely when $\Omega_E < -\frac{g_E}{n\psi k^{\frac{1}{n}} r_I^{\frac{n-1}{n}}}$ and $\Omega_I < -\frac{g_I}{n\psi k^{\frac{1}{n}} r_E^{\frac{n-1}{n}}}$, respectively (and so in

particular can only be negative if $\Omega_E < 0$ or $\Omega_I < 0$, respectively); (2) if $\Omega_E < 0$, then there is a stable fixed point with $r_E = 0$ at a finite positive value of c, which we calculate; and, (3) for $n = 2$, if in addition $\Omega_E < \frac{g_E^2}{g_I^2}\Omega_I$, then the set of fixed points $r_E(c)$ reaches a maximum for increasing c with $\frac{dr_E}{dc} = 0$ before being pushed to zero, and we calculate the corresponding c and peak value of r_E.

The condition $\Omega_E < 0$ states that the linear term in c in the high-c expansion for r_E (Eq. 15) is negative, driving r_E to zero; while the requirement in condition three states that the linear term in the expansion for r_I is either positive, or not so negative as to disrupt the ability of inhibition to drive r_E to zero.

These results suggest, but do not prove, that r_E will be driven to zero for arbitrary n whenever $\Omega_E < 0$ (although there is a stable fixed point with $r_E = 0$ at finite c, we have not proven that it is the only stable fixed point). We note that r_I can never be zero for finite c if $g_I > 0$ or $r_E > 0$, so even for $\Omega_I < 0$, r_I can never be driven to zero with increasing c.

For $\Omega_E < \Omega_I < 0$, $g_E = g_I$, we find in simulations that r_I only increases with increasing c, and that the values of c at which r_E goes to zero and, for $n = 2$, at which r_E peaks and the corresponding peak r_E all are as calculated (see Fig. 2A). We speculate that, for $\Omega_E < 0$ and $\Omega_E < \frac{g_E^{f(n)}}{g_I^{f(n)}}\Omega_I$, where $f(n)$ may equal n or may equal 2, r_E can never become large enough to set $\frac{dr_I}{dc} < 0$, while r_I always becomes large enough to set $\frac{dr_E}{dc} < 0$ and so ultimately to drive r_E to zero.

When $\Omega_I < \Omega_E < 0$ for $g_E = g_I$, we find unbiological behavior in simulations in which both r_E and r_I jump to very high levels at very low c, after which r_E monotonically decreases and is ultimately pushed to 0 (see Fig. 2E). Numerical calculations suggest a discontinuity at the jump, which may explain why our calculations do not find a zero of $\frac{dr_E}{dc}$ for real positive c in this case. We have not tried to analyze this behavior.

5.2.1 When can r_E or r_I decrease with contrast?

We define the matrix $\mathbf{\Phi_r} = nk^{\frac{1}{n}} \begin{pmatrix} r_E^{\frac{n-1}{n}} & 0 \\ 0 & r_I^{\frac{n-1}{n}} \end{pmatrix}$. Then a simple calculation shows that $\frac{d\mathbf{r}}{dc} = \mathbf{\Phi_r}(\psi \mathbf{J} \frac{d\mathbf{r}}{dc} + \mathbf{g})$ or $\frac{d\mathbf{r}}{dc} = (\mathbf{1} - \psi \mathbf{\Phi_r J})^{-1} \mathbf{\Phi_r g}$, which gives

$$\frac{d\mathbf{r}}{dc} = \frac{nk^{\frac{1}{n}} \begin{pmatrix} r_E^{\frac{n-1}{n}} \left(\Omega_E n \psi k^{\frac{1}{n}} r_I^{\frac{n-1}{n}} + g_E \right) \\ r_I^{\frac{n-1}{n}} \left(\Omega_I n \psi k^{\frac{1}{n}} r_E^{\frac{n-1}{n}} + g_I \right) \end{pmatrix}}{\text{Det } (\mathbf{1} - \psi \mathbf{\Phi_r J})} \tag{31}$$

Stability requires that Det $(\mathbf{1} - \mathbf{\Phi_r W}) > 0$. Thus, this expression shows that, for a stable fixed point, r_E or r_I decrease with contrast precisely when

$$\Omega_E < -\frac{g_E}{n\psi k^{\frac{1}{n}} r_I^{\frac{n-1}{n}}} \qquad \left(\frac{dr_E}{dc} < 0\right) \tag{32}$$

$$\Omega_I < -\frac{g_I}{n\psi k^{\frac{1}{n}} r_E^{\frac{n-1}{n}}} \qquad \left(\frac{dr_I}{dc} < 0\right) \tag{33}$$

5.2.2 The c at which r_E becomes 0

The $c > 0$ at which r_E first becomes 0 with increasing c can be determined as follows. First, at this c, $r_I = cg_E/\psi J_{EI}$, because this is the value of r_I that sets the input to r_E to zero when $r_E = 0$. The equation for the r_I steady state then yields $\frac{cg_E}{\psi J_{EI}} = k\left(cg_I - cg_E\frac{J_{II}}{J_{EI}}\right)^{.n} = kc^n\left(\frac{-\Omega_E}{J_{EI}}\right)^{.n}$. The right side gives zero unless $\Omega_E < 0$, so a solution for $c \neq 0$ exists only for $\Omega_E < 0$. In this case, one can solve to find $c = J_{EI}\left(\frac{g_E}{k\psi(-\Omega_E)^n}\right)^{\frac{1}{n-1}}$. This corresponds to $\alpha = \frac{J_{EI}^{n-1} g_E}{(-\Omega_E)^n}$ or $\beta = \frac{\Omega_E}{\left(J_{EI}^{n-1} g_E\right)^{\frac{1}{n}}}$.[9]

Note that any fixed point $y_E = 0$, $y_I > 0$ is stable for any q since the Jacobian matrix is $\mathcal{J} = n\alpha^{\frac{1}{n}}\begin{pmatrix} -1 & 0 \\ \frac{1}{q} y_I^{\frac{n-1}{n}} \psi J_{IE} & -\frac{1}{q}\left(y_I^{\frac{n-1}{n}} \psi J_{II} + 1\right) \end{pmatrix}$, which has two negative eigenvalues (equal to the two diagonal entries of $\mathcal{J}$).

This shows that $r_E = 0$, $r_I = cg_E/\psi J_{EI}$ is a stable fixed point for this value of c, but does not rule out the existence of other fixed points.

5.2.3 Peak firing rate and corresponding contrast

From the steady-state equation for r_E, we can solve for r_I in terms of r_E. Substituting this into the steady-state equation for r_I, we obtain an implicit equation for the steady-state firing rate of r_E:

$$r_E = \frac{1}{\psi \mathrm{Det}\,\mathbf{J}}\left(c\,\Omega_E - J_{II}\left(\frac{r_E}{k}\right)^{\frac{1}{n}} + J_{EI}\left(\frac{\psi J_{EE} r_E - \left(\frac{r_E}{k}\right)^{\frac{1}{n}} + cg_E}{\psi J_{EI} k}\right)^{\frac{1}{n}}\right) \tag{34}$$

We now restrict to the case $n = 2$. We take the derivative of both sides w.r.t. c, yielding an implicit equation for $\frac{dr_E}{dc}$ which we solve for $\frac{dr_E}{dc}$. We then solve for $\frac{dr_E}{dc} = 0$ to find the c at which

[9] Once r_E has been pushed to zero, for increasing c, r_I continues to increase according to $r_I = k(cg_I - \psi J_{II} r_I)^n$, which for $n = 2$ has the solution $r_I = \frac{\left(\sqrt{1+4cg_I\psi J_{II}k^2}-1\right)^2}{4k\psi^2 J_{II}^2}$, and r_E remains 0.

a maximum firing rate can occur:[10]

$$c^{\max} = \frac{g_E J_{EI}}{4\Omega_E^2 k\psi} + \frac{\sqrt{\frac{\tau_E}{k}} - J_{EE}\psi r_E}{g_E} \tag{35}$$

If we then substitute $c^{\max}$ for c in Eq. 34, we can solve explicitly for the corresponding maximum value of r_E, $r_E^{\max}$. For this to have a real value – that is, for a maximum of r_E as a function of c to exist – it must be the case that $\left(-\mathbf{J}^{-1}\mathbf{g}\right)_E < 0$. We restrict to the case $\mathrm{Det}\,\mathbf{J} > 0$, which is required for stability, so that the requirement for a real solution is $\Omega_E < 0$. We then obtain the following expression for $\mathbf{r}_E^{\max}$. With $\Omega_E < 0$, define:

$$\kappa = \frac{g_E^2\Omega_I + 2g_I^2|\Omega_E| - 2g_I\sqrt{|\Omega_E|\left(g_E^2\Omega_I + g_I^2|\Omega_E|\right)}}{|\Omega_E|\Omega_I^2} \tag{36}$$

Then the maximum firing rate r_E^{max} is[11]:

$$r_E^{\max} = \frac{\kappa}{4k\psi^2} \tag{37}$$

Note that for this to be real, an additional condition needs to be satisfied, namely $g_E^2\Omega_I > -g_I^2|\Omega_E|$.

Substituting this expression back into Eq. 35, we obtain an expression for $c^{\max}$ in terms of model parameters:

$$c^{max} = \frac{1}{4k\psi g_E}\left(\frac{g_E^2 J_{EI}}{\Omega_E^2} - J_{EE}\kappa + 2\sqrt{\kappa}\right) \tag{38}$$

Explicit calculation shows that this is always a maximum: the 2$^{\text{nd}}$ derivative $\frac{d^2 r_E}{dc^2} < 0$.[12] $c^{\max}$ is not guaranteed to be positive – this is governed by rather complicated conditions on the g's and J's – but in practice we have found it to be positive for the simulation parameters we have used.

When $c^{\max}$ is positive, then r_E reaches a maximum as a function of c at $c^{\max}$. Equations 37 and 38 show that, in this case, both the maximum excitatory firing rate that can be achieved by the network and the contrast at which this maximum is achieved decrease with increasing ψ. Thus, when the 2-D reduced model, Eqs. 20-21, accurately captures the high-dimensional model, Eqs. 18-19, then in the high-D model, if the stimulus is widened or a second stimulus is added, the maximum excitatory firing rate will go down and will occur at a lower contrast.

[10] These calculations were done in Mathematica. We solved for the numerator of the expression for $\frac{dr_E}{dc}$ being 0, assuming that the denominator was not simultaneously zero. We calculate further that when the numerator is zero ($c = c^{\max}$), and for $\Omega_E < 0$, the denominator is zero when $r_E = \frac{(g_I J_{EI})^2}{4k\psi^2(J_{EE}\Omega_E - g_E \mathrm{Det}\,\mathbf{J})^2}$. So long as this does not coincide with the value of r_E for $c = c^{\max}$ (Eq. 37) the procedure is fine.

[11] There is a second solution in which the sign in front of the square-root term in κ (Eq. 36) is positive. For parameters we have used in simulations the first solution gives positive c and the second gives negative c, so we have focused on the first solution, but there may be parameters for which this situation is reversed.

[12] For this 2nd solution mentioned in footnote 11, the 2nd derivative is positive precisely when $\Omega_I > 0$.

$c^{\max}$ and $\mathbf{r}^{\max}$ correspond to $\alpha = \frac{1}{4g_E}\left(\frac{g_E^2 J_{EI}}{\Omega_E^2} - J_{EE}\kappa + 2\sqrt{\kappa}\right)$ and $y_E = \frac{\psi}{c^{\max}} r_E^{\max} = \frac{\kappa g_E}{\left(\frac{g_E^2 J_{EI}}{\Omega_E^2} - J_{EE}\kappa + 2\sqrt{\kappa}\right)}$. However, note that this is not a maximum of the y_E vs. α curve, but rather occurs for α higher than that peak, where the curve has a negative slope. We saw in section 4.2 that $\frac{d\mathbf{r}}{dc} = \frac{(n-1)\alpha}{\psi}\left(\frac{d\mathbf{y}}{d\alpha} + \frac{\mathbf{y}}{(n-1)\alpha}\right)$, so $\frac{d\mathbf{r}}{dc} = 0$ implies $\frac{d\mathbf{y}}{d\alpha} = -\frac{\mathbf{y}}{(n-1)\alpha}$, *i.e.* $\mathbf{y}$ is locally evolving as $\alpha^{-(n-1)}$.

In sum, for $\mathrm{Det}\,\mathbf{J} > 0$ and $n = 2$, the steady state solution for r_E has a maximum value as a function of c, given by Eq. 37, precisely when (1) $\Omega_E < 0$ and (2) $\Omega_E < \frac{g_E^2}{g_I^2}\Omega_I$.

5.3 Steady-state solutions for different parameter regimes

In Fig. 2 we illustrate model behavior, as a function of stimulus strength c, for 5 parameter regimes, with $\mathrm{Det}\,\mathbf{J} > 0$, $n = 2$, and $g_E = g_I$ in all cases. The 5 parameter regimes are: $\Omega_E < 0$ and $\Omega_I < 0$, with either $\Omega_E < \Omega_I$ (Fig. 2A) or $\Omega_E > \Omega_I$ (Fig. 2E); $\Omega_E < 0$ and $\Omega_I > 0$ (Fig. 2B); and $\Omega_E > 0$ and $\Omega_I > 0$, with either $\Omega_E < \Omega_I$ (Fig. 2C) or $\Omega_E > \Omega_I$ (Fig. 2D). We chose parameters relatively arbitrarily, by starting with a set of parameters that had worked well in simulations of the ring model (Column A) and changing small sets of parameters to change the regime. However in small amounts of studies of other parameters in the different regimes we have found behaviors to be similar, with one exception. For the case $\Omega_E < 0$ and $\Omega_I > 0$ (column B), the excitatory firing rate can peak and be driven to zero without ever reaching a level at which the excitatory subnetwork is unstable, as manifest in the figures as stability for all possible values of q (bottom row, described below). This occurs because $\Omega_I > 0$ is compatible with $J_{EE} = 0$, and so for weak enough J_{EE} the behavior can be similar to $J_{EE} = 0$ behavior. (For $\Omega_E > 0$ and $\Omega_I > 0$, r_E ultimately grows linearly with increasing c for large enough c, so given any small but finite J_{EE} the excitatory subnetwork must ultimately reach instability.)

We illustrate behavior across a large range of c, sufficient to include the point at which r_E is pushed to zero for cases with $\Omega_E < 0$. However, we imagine the dynamic range of cortex, corresponding to the dynamic range of the firing rates of the inputs to cortex, corresponds to a smaller range, *e.g.* through $c = 100$ (as in Fig. 1A, reduced model, 1 stimulus, which uses essentially the same parameters as Fig. 2A).

For each set of parameters, we first illustrate firing rates (top row), with red and blue indicating r_E and r_I respectively. As expected, parameters with $\Omega_E < 0$ (columns A,B,E) all show r_E eventually pushed to zero with increasing c, while those with $\Omega_E > 0$ (columns C,D) show r_E moving toward linear growth with increasing c. The combination $\Omega_I < \Omega_E < 0$ (column E) leads, as mentioned previously, to unbiological behavior in which both E and I rates abruptly jump (discontinuously, in numerical calculations with c discretized in 0.00001 steps) to high rates at low c, after which r_E monotonically falls with increasing c.

We next illustrate normalization weights (second row), computed just as in Fig. 1, so that weights > 1 indicate supralinear summation in the corresponding ring model and weights < 1 indicate sublinear summation. All but the case $\Omega_E > \Omega_I > 0$ show a regime of supralinear summation for very low contrasts (behavior in all cases is sublinear for $c > 10$), although the

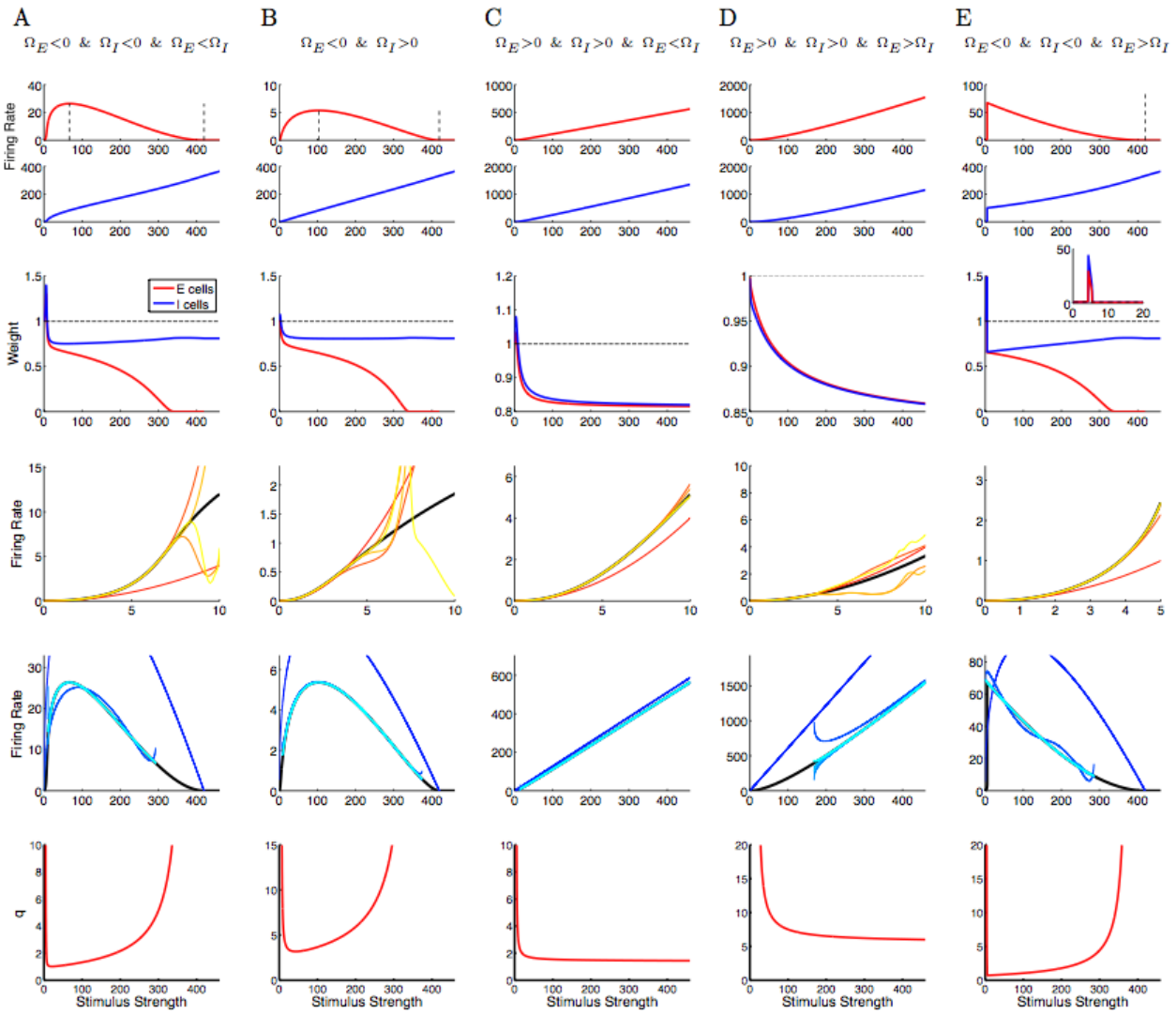

Figure 2:

Behavior of the 2D Model in Different Parameter Regimes. Each column indicates a different parameter set, as indicated. In all cases, $\text{Det}\,\mathbf{J} > 0$, $n = 2$ and $g_E = g_I = \frac{1}{\sqrt{2}}$. The first column uses the same parameters as the 2-D reduced model in Fig. 1, except that weights here were rounded to 2 significant figures. In all figures the horizontal axis is stimulus strength c. **Top row**: E (red) and I (blue) firing rates, r_E and r_I, at fixed point. For cases with $\Omega_E < 0$, dashed vertical lines indicate analytic prediction for c at which r_E peaks (Eq. 38) and at which r_E goes to zero (Section 5.2.2). **Second Row**: Weights reflecting supralinear summation (weight > 1) or sublinear summation (weight < 1) in an equivalent ring model, computed as in Fig. 1B. Again, red and blue indicate E- and I-cells, respectively. Inset in column E shows supralinear responses at low values of c. **Third Row**: Iterative solutions for r_E in the low-contrast regime. We use (Eq. 39) for y_E, and graph $r_E[t] = y_E[t]c/\psi$ vs. c, where t is number of iterations. Black curves are exact solutions; red to yellow represent iterative solutions for 1, 5, 10, 14, 19 iterations. Iterative solutions are shown only over the range for which they are real. **Fourth Row**: Iterative solutions for r_E in the high-contrast regime, using Eq. 40. Conventions as in 3rd row, except now blue to cyan represent 1, 5, 10, 14, 19 iterations. **Fifth Row**: Values of $q = \tau_I/\tau_E$ separating regions in which fixed point is stable (below red line) vs. unstable (above red line). Parameters used: $\mathbf{W} = \psi\mathbf{J}$, where $\psi = 54.86$, the value of Ψ in Fig. 1 for the default Gaussian width. The elements of $\mathbf{J}$ are: $J_{EE} = 0.044$ except in (D), where $J_{EE} = 0.014$; $J_{EI} = 0.023$; (A) $J_{IE} = 0.042$; $J_{II} = 0.018$; (B) $J_{IE} = 0.082$; $J_{II} = 0.018$; (C) $J_{IE} = 0.082$; $J_{II} = 0.038$; (D) $J_{IE} = 0.062$; $J_{II} = 0.088$; (E) $J_{IE} = 0.039$; $J_{II} = 0.018$. We convert from α to c using $k = 0.04$. This yields the following values of α for columns A to E, respectively: $\alpha = .148c$, $.213c$, $.226c$, $.243c$, and $.144c$.

supralinear behavior is weak for $\Omega_I > 0$.

The third and fourth rows of Fig. 2 illustrate the iterative solutions that stem from the scaling solutions in the low- and high-contrast regimes. The values of $\mathbf{J}$ used (listed in legend of Fig. 2) are not normalized to have $\|\mathbf{J}\| = 1$, so for these iterations we take $\hat{\mathbf{J}} = \mathbf{J}/\|\mathbf{J}\|$ and $\hat{\psi} = \psi\|\mathbf{J}\|$ where $\|\mathbf{J}\|$ is the 2-norm of $\mathbf{J}$ (the maximum singular value of $\mathbf{J}$), with $\alpha = kc^{n-1}\hat{\psi}$. We then show results as $r_E = y_E c/\hat{\psi}$ vs. c.

The small-α (low contrast) iterations are shown in the third row of Fig. 2. Here, we treat the equation $\mathbf{y} = \alpha(\hat{\mathbf{J}}\mathbf{y} + \mathbf{g})^{\cdot n}$ as the recurrence relation

$$\mathbf{y}[t] = \alpha(\hat{\mathbf{J}}\mathbf{y}[t-1] + \mathbf{g})^{\cdot n} \tag{39}$$

with the starting condition $y[0] = 0$. Results are shown for numbers of iterations ranging from 1 to 19. As few as 5 iterations gives a good approximation for small c, while increasing the number of iterations to 19 adds little. The low-contrast iterations all fail before or very slightly after $c = 10$, which corresponds to α in the range 1.4 to 2.4 across the parameters. That is, the failure occurs for $\alpha = O(1)$, as expected.

For the high-α or small-β (high contrast) case, we treat the equation $\mathbf{y} = \hat{\mathbf{J}}^{-1}\left(-\mathbf{g} + \beta\mathbf{y}^{\cdot\frac{1}{n}}\right)$ as the recurrence relation:

$$\mathbf{y}[t] = \hat{\mathbf{J}}^{-1}\left(-\mathbf{g} + \beta\mathbf{y}[t-1]^{\cdot\frac{1}{n}}\right) \tag{40}$$

We use as starting conditions $y_E[0] = 0$ with $y_I[0] = 0$ for $\Omega_E > 0$, $y_I[0] = g_I/\hat{J}_{EI}$ for $\Omega_E < 0$. For $\Omega_E < 0$, using $\mathbf{y}[0] = 0$ would give complex solutions. We instead use as a starting condition the value of $\mathbf{y}$ when y_E reaches zero with increasing c. The fourth row of Fig. 2 illustrates these high-contrast solutions. Again, 5 iterations do about as well as larger numbers of iterations. The iterations give good approximations for high c but, for $\Omega_E < 0$, fail for larger c as r_E approaches zero. β is very small for these large c's, so this presumably represents the initial conditions no longer being in the basin of attraction of the fixed point. For low c failure of convergence is expected for $\beta = O(1)$ (recall that β increases with decreasing c, $\beta = 1/\sqrt{\alpha}$), although problems with the basin of attraction could also arise. None of the iterations work for c below about 9 or 10, corresponding to β roughly in the range .65 to .85, with the exception of column E. In that column, the largest number of iterations works down to the jump in r_E and r_I, which occurs at about $c = 5.435$ for the given parameters, or β around 1.1. In column D the iterations do not work below c about 190, which corresponds to β about 0.15, a bit lower than expected.

Finally, in the fifth row of Fig. 2, we show the value of $q = \frac{\tau_I}{\tau_E}$ that divides stability (values below curves) from instability (values above curves) of the fixed point, according to Eq. 29. In all cases except $\Omega_I < \Omega_E < 0$, the fixed point remains stable for $q < 1$ across the range of studied stimulus strengths, indicating that fine tuning or unreasonably small values of q are not required.

6 Discussion

We have shown in studies of a 2-D system (and found in simulation studies of higher-dimensional systems, to be presented elsewhere) that the supralinear network will dynamically stabilize with

increasing input strength provided the $I \Rightarrow E$ and $E \Rightarrow I$ connections mediating feedback inhibition are sufficiently strong and the inhibitory time constant is not too slow. This dynamic stabilization results in a change from responses scaling supralinearly to responses scaling sublinearly with input strength. The system can also yield "supersaturation", in which excitatory firing rates reach a peak with increasing input strengths and then decrease (as observed biologically, Ledgeway et al. 2005, Li and Creutzfeldt 1984, Peirce 2007, Tyler and Apkarian 1985), with rates ultimately decreasing to zero for large enough input strengths (which presumably are beyond the dynamical range of biological inputs). The conditions for this to occur were characterized in the 2-D system. The strongest sublinear behavior, and hence behavior most likely to underly biological observations in cerebral cortex, occurs for parameters that lead to supersaturation.

Many questions remain outstanding. As some examples: within the range of models analyzed here, can more precise results, analogous to those obtained here for 2-dimensional models, be obtained for higher-dimensional models, for which we only discussed general scaling arguments? For any dimensionality, can useful results be obtained as to when the network is globally stable? How will diversity of network parameters, including in particular of the power n, alter behavior? Presumably an even slightly larger mean n for I vs. E cells will enormously enhance the range of parameters that will stabilize. How will cell-to-cell variability of n affect behavior? How will behavior be affected by taking into account the decreased noise level, and thus increase in n, that occurs with increasing stimulus contrast (Finn et al. 2007), *i.e.* with increasing input firing rate? How will network behavior be modified by addition of short-term synaptic facilitation and depression (*e.g.*, Fioravante and Regehr 2011)? Can analysis be done of more biophysically realistic models, such as networks of integrate-and-fire neurons, which have an input/output function well approximated by a power law so long as they are firing on input fluctuations rather than the mean input (Hansel and van Vreeswijk 2002)? What can we learn as we move beyond the steady state to network dynamics, particularly using more realistic models that can better capture faster dynamics and that incorporate synaptic delays? How will the network behave when multiple types of inhibitory neurons (*e.g.* Isaacson and Scanziani 2011), or of excitatory neurons, are incorporated? Incorporating multiple neuronal subtypes must presumably be guided by knowledge not yet available of the separate connectivity patterns as well as biophysical properties of the different subtypes.

Despite the many open questions, we believe the basic findings are likely to be quite robust and to underly a wide range of cerebral cortical behavior: networks with supralinear input/output functions can dynamically stabilize, resulting in a transition from supralinear to sublinear input summation.

References

J. S. Anderson, M. Carandini, and D. Ferster. Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J. Neurophysiol.*, 84:909–926, 2000a.

J. S. Anderson, I. Lampl, D. Gillespie, and D. Ferster. The contribution of noise to contrast invariance of orientation tuning in cat visual cortex. *Science*, 290:1968–1972, 2000b.

J. S. Anderson, I. Lampl, D. C. Gillespie, and D. Ferster. Membrane potential and conductance changes underlying length tuning of cells in cat primary visual cortex. *J. Neurosci.*, 21:2104–2112, 2001.

R. M. Bruno. Synchrony in sensation. *Curr. Opin. Neurobiol.*, 21:701–708, 2011.

M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, 13:51–62, 2011.

J. R. Cavanaugh, W. Bair, and J. A. Movshon. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.*, 88:2530–2546, 2002.

P. Dayan and L. F. Abbott. *Theoretical Neuroscience*. MIT Press, Cambridge, MA, 2001.

G. B. Ermentrout and D. H. Terman. *Mathematical Foundations of Neuroscience*. Springer, New York, 2010.

D. Ferster. Orientation selectivity of synaptic potentials in neurons of cat primary visual cortex. *J. Neurosci.*, 6:1284–1301, 1986.

D. Ferster and K. D. Miller. Neural mechanisms of orientation selectivity in the visual cortex. *Ann. Rev. Neurosci.*, 23:441–471, 2000.

I. M. Finn, N. J. Priebe, and D. Ferster. The emergence of contrast-invariant orientation tuning in simple cells of cat visual cortex. *Neuron*, 54:137–152, 2007.

D. Fioravante and W. G. Regehr. Short-term forms of presynaptic plasticity. *Curr. Opin. Neurobiol.*, 21:269–274, 2011.

W. Gerstner and W. Kistler. *Spiking Neuron Models*. Cambridge University Press, Cambridge, UK, 2002.

D. Hansel and C. van Vreeswijk. How noise contributes to contrast invariance of orientation tuning in cat visual cortex. *J. Neurosci.*, 22:5118–5128, 2002.

H. W. Heuer and K. H. Britten. Contrast dependence of response normalization in area MT of the rhesus macaque. *J. Neurophysiol.*, 88:3398–3408, Dec 2002.

J. S. Isaacson and M. Scanziani. How inhibition shapes cortical activity. *Neuron*, 72:231–243, 2011.

T. Z. Lauritzen, A. E. Krukowski, and K. D. Miller. Local correlation-based circuitry can account for responses to multi-grating stimuli in a model of cat V1. *J. Neurophysiol.*, 86:1803–1815, 2001.

T. Ledgeway, C. Zhan, A. P. Johnson, Y. Song, and C. L. Baker. The direction-selective contrast response of area 18 neurons is different for first- and second-order motion. *Vis. Neurosci.*, 22: 87–99, 2005.

B. Li, J. K. Thompson, T. Duong, M. R. Peterson, and R. D. Freeman. Origins of cross-orientation suppression in the visual cortex. *J. Neurophysiol.*, 96:1755–1764, Oct 2006.

C. Y. Li and O. Creutzfeldt. The representation of contrast and other stimulus parameters by single neurons in area 17 of the cat. *Pflugers. Arch.*, 401:304–314, 1984.

M. London, A. Roth, L. Beeren, M. Hausser, and P. E. Latham. Sensitivity to perturbations in vivo implies high noise and suggests rate coding in cortex. *Nature*, 466:123–127, 2010.

J. Marino, J. Schummers, D. C. Lyon, L. Schwabe, O. Beck, P. Wiesing, K. Obermayer, and M. Sur. Invariant computations in local cortical networks with balanced excitation and inhibition. *Nature Neurosci.*, 8:194–201, 2005.

L.M. Martinez, J.M. Alonso, R.C. Reid, and J.A. Hirsch. Laminar processing of stimulus orientation in cat visual cortex. *J. Physiol.*, 540:321–33, 2002.

K. D. Miller and F. Fumarola. Mathematical equivalence of two common forms of firing rate models of neural networks. *Neural Comput*, 24:25–31, 2012.

K. D. Miller and D. B. Rubin. Contrast dependence of summation field size and surround properties in a nonlinear circuit model of V1. *Program No. 126.2. 2010 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience*, Online, 2010.

K. D. Miller and D. B. Rubin. Balanced amplification and normalization in a simple circuit model of visual cortex explain multiple aspects of attentional modulation. *Program No. 428.09. 2011 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience*, Online, 2011.

K. D. Miller and T. W. Troyer. Neural noise can explain expansive, power-law nonlinearities in neural response functions. *J. Neurophysiol.*, 87:653–659, 2002.

T. Ohshiro, D. E. Angelaki, and G. C. DeAngelis. A normalization model of multisensory integration. *Nat. Neurosci.*, 14:775–782, 2011.

I. Ohzawa, G. Sclar, and R. D. Freeman. Contrast gain control in the cat's visual system. *J. Neurophysiol.*, 54:651–667, 1985.

H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. Inhibitory stabilization of the cortical network underlies visual surround suppression. *Neuron*, 62:578–592, 2009.

J. W. Peirce. The potential importance of saturating and supersaturating contrast response functions in visual cortex. *J Vis*, 7:13, 2007.

U. Polat, K. Mizobe, M. W. Pettet, T. Kasamatsu, and A. M. Norcia. Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, 391:580–584, 1998.

N. J. Priebe and D. Ferster. Direction selectivity of excitation and inhibition in simple cells of the cat primary visual cortex. *Neuron*, 45:133–45, 2005.

N. J. Priebe and D. Ferster. Mechanisms underlying cross-orientation supression in cat visual cortex. *Nature Neurosci.*, 9:552–561, 2006.

N.J. Priebe, F. Mechler, M. Carandini, and D. Ferster. The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nat. Neurosci.*, 7(10):1113–22, 2004.

A. Renart, J. de la Rocha, P. Bartho, L. Hollender, N. Parga, A. Reyes, and K. D. Harris. The asynchronous state in cortical circuits. *Science*, 327:587–590, 2010.

D. B. Rubin and K. D. Miller. Normalization in a nonlinear circuit model of V1. *Program No. 126.1. 2010 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience*, Online, 2010.

D. B. Rubin and K. D. Miller. Normalization in a simple circuit model of visual cortex explains stimulus-induced reduction in shared variability. *Program No. 428.10. 2011 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience*, Online, 2011.

M.P. Sceniak, D. L. Ringach, M.J. Hawken, and R. Shapley. Contrast's effect on spatial summation by macaque v1 neurons. *Nature Neurosci.*, 2:733–739, 1999.

F. Sengpiel, C. Blakemore, and A. Sen. Characteristics of surround inhibition in cat area 17. *Exp. Brain Res.*, 116:216–228, 1997.

S. Shushruth, J. M. Ichida, J. B. Levitt, and A. Angelucci. Comparison of spatial summation properties of neurons in macaque V1 and V2. *J. Neurophysiol.*, 102:2069–2083, Oct 2009.

B. C. Skottun, A. Bradley, G. Sclar, I. Ohzawa, and R. D. Freeman. The effects of contrast on visual orientation and spatial frequency discrimination: A comparison of single cells and behavior. *J. Neurophysiol.*, 57:773–786, 1987.

X. M. Song and C. Y. Li. Contrast-dependent and contrast-independent spatial summation of primary visual cortical neurons of the cat. *Cerebral Cortex*, 18:331–336, 2008.

M. V. Tsodyks, W. E. Skaggs, and B. L. Sejnowski, T. J.and McNaughton. Paradoxical effects of external modulation of inhibitory interneurons. *J. Neurosci.*, 17:4382–4388, 1997.

J. M. Tsui and C. C. Pack. Contrast sensitivity of MT receptive field centers and surrounds. *J. Neurophysiol.*, 106:1888–1900, 2011.

C. W. Tyler and P. A. Apkarian. Effects of contrast, orientation and binocularity in the pattern evoked potential. *Vision Res.*, 25:755–766, 1985.

C. van Vreeswijk and H. Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274:1724–1726, 1996.

C. van Vreeswijk and H. Sompolinsky. Chaotic balanced state in a model of cortical circuits. *Neural Computation*, 10:1321–1371, 1998.